13 May 2022

Dog10K Consortium Data Release

The Dog10K consortium is an international collaborative effort to sequence the genomes of 10,000 canids from a variety of diverse breeds and populations. This release of high confidence variants from 1929 samples marks the largest publicly available whole genome sequence canine genotype catalogue to date. Consistent with the goals of the Dog10K consortium, the dataset is designed to be useful to those investigating the evolutionary history of dogs, genetic changes associated with domestication, and relationships of genotype to phenotype. Several Dog10K working groups have been established to characterize and investigate the dataset with the intention of publishing our findings related to genome-wide patterns of variation, breed ancestry and relationships, signals of selection, mitochondrial diversity, structural variant annotation, and the impact of variation on gene function.

Below, we have provided a summary of methods used to generate this dataset and an acknowledgements statement that users of these data should include in any publication using data both prior to publication of the primary paper and subsequently.

Dog10K Data Use Acknowledgement

Genomic variation data was produced by the Dog10K Project, an international collaboration to advance canine genetics. Genome sequencing was supported by National Science and Technology Innovation 2030 Major Project of China (2021ZD0203900) and the National Key R&D Program of China (2019YFA0707101).

Methods

Phase I Dog10K samples were paired-end whole genome sequenced to approximately 20x coverage with a mean insert length of either 300 or 350 bp. Sequencing was performed on an Illumina HiSeq X Ten instrument. Sample data was then processed using GATK version 4.2.0.0 according to the steps outlined in https://github.com/jmkidd/dogmap/blob/main/README.md. The reference assembly used for sample read alignment and variant databases used for base quality score recalibration (BQSR) and variant quality score recalibration (VQSR) can be found at https://kiddlabshare.med.umich.edu/public-data/UU_Cfam_GSD_1.0-Y/. To ensure ploidy of sex chromosomes was consistent with sample sex, genotypes from the non-pseudoautosomal region (PAR) of chromosome X (chrX_non-PAR) were called separately from the combined set of autosomes and chromosome X PAR (chrX:1-6605250) (Auto+PAR). For chrX_non-PAR, males were treated as haploid. Although assembled scaffolds for chromosome Y were included in the reference, these regions were not genotyped in phase I of Dog10K. After genotype calling, samples and variants underwent various quality control filtering procedures. In brief, initial sample quality filtering was performed based on site specific measurements taken at Illumina CanineHD BeadChip array positions according to the following criteria: low coverage, skewed reference read fraction, sample duplicate, and mislabeled dog breed. Next, variants were filtered as SNVs and non-SNVs and separated according to chromosomal region (Auto+PAR and chrX_non-PAR). Auto+PAR and chrX_non-PAR SNVs underwent VQSR using the

Illumina CanineHD BeadChip and Axiom Canine HD Genotyping Array positions as a truth set. SNVs within the 99.0% truth sensitivity tranche were declared as PASS SNVs. Due to the lack of a truth set for non-SNVs, the following thresholds were used to identify low quality non-SNV variants, QD < 2.0, FS > 200.0, ReadPosRankSum < -2.0, and SOR > 10.0. After variant quality control (QC), two additional rounds of sample quality control filtering took place. The first round of QC filtering was based on excessive per sample TiTv ratios, singleton counts, and missingness counts (five standard deviations above mean). The second and final round of sample QC was based on genetic similarity between samples to validate each sample's breed label and population label. A neighbor joining tree of all remaining Dog10K breed samples was constructed. Dog10K breed labels were validated according to whether samples clustered with members of the same or similar breeds. Samples whose clustering pattern was inconsistent with their breed label were reviewed and updated according to the submitting intuition's metadata. Inconsistent samples with no additional evidence to suggest a corrected breed label were removed from the analysis.

The posted SNV VCF file contains variants from 1,987 samples. The additional analyses described above identified a subset of 1,929 samples with high variant call metrics. We recommend that analyses be limited to variants founds in the 1,929 samples. Quality control of structural variant analyses identified additional samples with poor sequence quality metrics. As a result, structural variant analyses in the Dog10K consortium is limited to a subset of 1,879 samples. Sample information, including inclusion status in these different sample sets, is described in the file dog10K-alignment-sample-table.2022-02-23-v7.xlsx. The files described above are available at https://kiddlabshare.med.umich.edu/dog10K/ and will be submitted to appropriate databases upon manuscript submission.